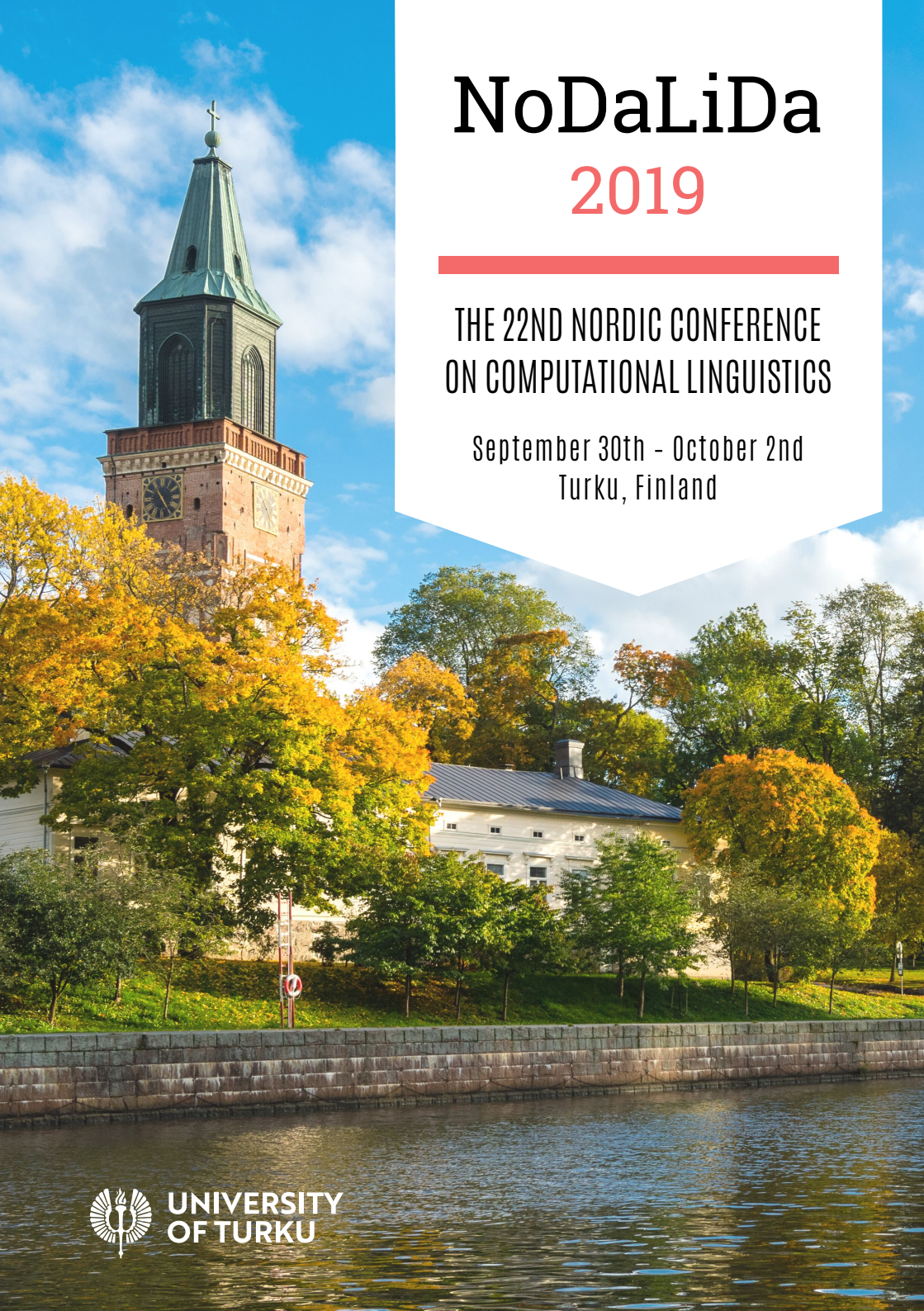


NoDaLiDa 2019

THE 22ND NORDIC CONFERENCE ON COMPUTATIONAL LINGUISTICS

September 30th - October 2nd
Turku, Finland

UNIVERSITY OF TURKU

Cover photo: © The City of Turku, Seilo Ristimäki

Welcome

Welcome to the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa 2019) held at the University of Turku in the beautiful city of Turku in Finland, on September 30-October 2, 2019. The aim of NoDaLiDa is to bring together researchers in the Nordic countries interested in any aspect related to human language and speech technologies. It is a great honor for me to serve as the general chair of NoDaLiDa 2019.

NoDaLiDa has a very long tradition. It stems from a working group initiative led by Sture Allèn, Kolbjörn Heggstad, Baldur Jönsson, Viljo Kohonen and Bente Maegaard (as the preface of the oldest workshop proceedings in the ACL anthology reveals).[1] They organized the first NoDaLiDa ("Nordiska datalingvistikdagar") in Gothenburg on October 10-11, 1977. In 2006, NEALT, the Northern European Association for Language Technology was founded. We are very honored to bring this bi-annual conference after 42 years to Turku this fall.

We solicited three different types of papers (long, short, demo papers) and received 78 valid submissions. In total, we accepted 50 papers, which will be presented as 34 oral presentations, 10 posters and 5 papers. A total of 4 submissions were withdrawn in the process. Each paper was reviewed by three experts. We are extremely grateful to the Programme Committee members for their detailed and helpful reviews. Overall, there are 10 oral sessions with talks and one poster session organized into themes over the two days, starting each day with a keynote talk.

We would like to thank our two keynote speakers for travel to Turku and sharing their work. Marie-Catherine de Marneffe from Ohio State University will talk about "Do you know that there's still a chance? Identifying speaker commitment for natural language understanding". Grzegorz Chrupała from Tilburg University will talk about "Investigating neural representations of speech and language". We are also very grateful to Fred Karlsson, who accepted to share his insights into the Finnish language in the traditional NoDaLiDa language tutorial.

The conference is preceded by 5 workshops on a diverse set of topics: deep learning for natural language processing, NLP for Computer-Assisted Language Learning, Constraint Grammar Methods, Tools and Applications, NLP and pseudonymisation and Financial Narrative Processing. This shows the breadth of topics that can be found in language technology these days, and we are extremely happy and grateful to the workshop organizers for complementing the main program this way.

There will be two social events. A reception which is sponsored by the City of Turku and held at the Old Town Hall in Turku. A conference dinner will be held in the Turku Castle in the King's hall. Two fantastic evenings are awaiting.

[1] https://www.aclweb.org/anthology/events/ws-1977/

I would like to thank the entire team that made NoDaLiDa 2019 possible in the first place. First of all, I would like to thank Beáta Megyesi for inviting me to take up this exciting (and admittedly at times demanding) role and all her valuable input regarding NEALT and previous editions of NoDaLiDa. Jörg Tiedemann, for the smooth transition from the previous NoDaLiDa edition and his input and work as program chair; the program chair committee Jurgita Kapočiūtė-Dzikienė, Hrafn Loftsson, Patrizia Paggio, and Erik Velldal, for working hard on putting the program together. I am particularly grateful to Jörg Tiedemann, Jurgita Kapočiūtė-Dzikienė, Kairit Sirts and Patrizia Paggio for leading the reviewing process. Special thanks goes to the workshop chairs Richard Johansson and Kairit Sirts, who have done an invaluable job with leading the workshop selection and organization. A big thanks also to Miryam de Lhoneux for her work as social media chair and Mareike Hartmann for leading the publication efforts that led to this volume, as well as the coordination of the workshop proceedings. Thank you! Finally, my ultimate thanks goes to the amazing local organization committee and team. Thank you, Filip Ginter and Jenna Kanerva. With your infinite support and pro-active engagement in organizing NoDaLiDa you are the ones that make NoDaLiDa possible and surely an unforgettable experience. Thanks also to the entire local team (with special thanks to Hans Moen for help with the program): Li-Hsin Chang, Rami Ilo, Suwisa Kaewphan, Kai Hakala, Roosa Kyllönen, Veronika Laippala, Akseli Leino, Juhani Luotolahti, Farrokh Mehryary, Hans Moen, Maria Pyykönen, Sampo Pyysalo, Samuel Rönnqvist, Antti Saloranta, Antti Virtanen, Sanna Volanen. NoDaLiDa 2019 has received financial support from our generous sponsors, which we would also like to thank here.

Danke - kiitos!

We very much hope that you will have an enjoyable and inspiring time at NoDaLiDa 2019 in Turku.

Barbara Plank
København
September 2019

This is the usual place for the greetings from the local organizers, but as we set out to write it, it turns out that Barbara already said it all. So we really only need to add one thing: huge thanks to Barbara for all the hard work she put into NoDaLiDa. We can only wonder where you found the time for all this. We hope the Turku edition of NoDaLiDa will be a success, at least we tried our best to make it so.

Filip, Jenna, and the local team

Contents

Sponsors


Kindly.ai


NEALT
Northern European Association for Language Technology


TURKU
ÅBO

Organization

General Chair

Barbara Plank, IT University of Copenhagen, Denmark

Program Chairs

Jurgita Kapočiūtė-Dzikienė, Vytautas Magnus University, Lithuania
Hrafn Loftsson, Reykjavík University, Iceland
Patrizia Paggio, University of Copenhagen, Denmark
Jörg Tiedemann, University of Helsinki, Finland
Erik Velldal, University of Oslo, Norway

Publication Chair

Mareike Hartmann, University of Copenhagen, Denmark

Social Media Chair

Miryam de Lhoneux, Uppsala University, Sweden

Workshop Chairs

Richard Johansson, Chalmers Technical University and University of Gothenburg, Sweden
Kairit Sirts, University of Tartu, Estonia

Local Organizers

Filip Ginter (chair)
Jenna Kanerva (chair)
Li-Hsin Chang
Rami Ilo
Suwisa Kaewphan
Kai Hakala
Veronika Laippala
Akseli Leino
Juhani Luotolahti
Farrokh Mehryary
Hans Moen
Maria Pyykönen
Sampo Pyysalo
Samuel Rönnqvist
Antti Saloranta
Antti Virtanen
Sanna Volanen

Program at a glance

Monday, September 30

08:00-	Registration opens
09:00-17:00	NLPL Workshop on Deep Learning for Natural Language Processing
09:00-17:30	NLP for Computer-Assisted Language Learning (NLP4CALL)
09:00-15:30	Constraint Grammar - Methods, Tools and Applications
14:00-17:00	Workshop on NLP and Pseudonymisation
09:00-15:30	Financial Narrative Processing Workshop (FNP 2019)
19:00	Welcome Reception in the Turku City Hall

Tuesday, October 1

09:00-09:15	Opening
09:15-10:05	Keynote by Marie-Catherine de Marneffe
10:05-10:35	Coffee break
10:35-12:15	Parallel sessions
12:15-13:45	Lunch break
13:45-15:00	Parallel sessions
15:00-15:30	Coffee Break
15:30-16:45	Parallel sessions
16:45-17:45	Poster and demo session
19:30-23:00	Conference Dinner in the Turku Castle

Wednesday, October 2

09:00-09:50	Keynote by Grzegorz Chrupała
09:50-10:20	Coffee break
10:20-12:00	Parallel sessions
12:00-13:00	Lunch break
13:00-14:00	NEALT business meeting
14:00-15:15	Parallel sessions
15:15-15:45	Coffee Break
15:45-16:25	Tutorial on Finnish by Fred Karlsson
16:25-16:35	Closing and announcement of NoDaLiDa'21

Detailed program

Monday September 30, 2019: Workshops

Time	Session	Room
08:00-09:00	Registration	
	Workshops	
09:00-17:00	NLPL Workshop on Deep Learning for Natural Language Processing	PUB2
09:00-17:30	NLP for Computer-Assisted Language Learning (NLP4CALL)	PUB5
09:00-15:30	Constraint Grammar - Methods, Tools and Applications	PUB 209
14:00-17:00	Workshop on NLP and Pseudonymisation	PUB4
09:00-15:30	Financial Narrative Processing Workshop (FNP 2019)	PUB 126
10:00-10:30	Coffee break	
12:00-14:00	Lunch break	Holiday Club Caribia
15:00-15:30	Coffee break	
19:00	Welcome Reception in the Turku City Hall	

Tuesday October 1, 2019

Time	Session	Room
09:00-09:15	Opening	PUB1
09:15-10:05	***Keynote by Marie-Catherine de Marneffe:* Do you know that there's still a chance? Identifying speaker commitment for natural language understanding** *Chair: Joakim Nivre*	PUB1
10:05-10:35	Coffee break	
10:35-12:15	**Parallel session A: Multilinguality and Machine Translation** *Chair: Jörg Tiedemann*	PUB1
10:35-11:00	Comparison between NMT and PBSMT Performance for Translating Noisy User-Generated Content *José Carlos Rosales Nuñez, Djamé Seddah and Guillaume Wisniewski*	
11:00-11:25	Bootstrapping UD treebanks for Delexicalized Parsing *Prasanth Kolachina and Aarne Ranta*	

11:25-11:50	Lexical Resources for Low-Resource PoS Tagging in Neural Times *Barbara Plank and Sigrid Klerke*	
11:50-12:15	Toward Multilingual Identification of Online Registers *Veronika Laippala, Roosa Kyllönen, Jesse Egbert, Douglas Biber and Sampo Pyysalo*	
10:35-12:15	**Parallel session B: Embeddings, Biases and Language Change** *Chair: Richard Johansson*	PUB3
10:35-11:00	Gender Bias in Pretrained Swedish Embeddings *Magnus Sahlgren and Fredrik Olsson*	
11:00-11:25	A larger-scale evaluation resource of terms and their shift direction for diachronic lexical semantics *Astrid van Aggelen, Antske Fokkens, Laura Hollink and Jacco van Ossenbruggen*	
11:25-11:50	Some steps towards the generation of diachronic WordNets *Yuri Bizzoni, Marius Mosbach, Deitrich Klakow and Stefania Degaetano-Ortlieb*	
11:50-12:15	An evaluation of Czech word embeddings *Karolína Hořeňovská*	
12:15-13:45	Lunch break	Holiday Club Caribia
13:45-15:00	**Parallel session A: Semantics** *Chair: Marianna Apidianaki*	PUB1
13:45-14:10	Language Modeling with Syntactic and Semantic Representation for Sentence Acceptability Predictions *Adam Ek, Jean-Philippe Bernardy and Shalom Lappin*	
14:10-14:35	Comparing linear and neural models for competitive MWE identification *Hazem Al Saied, Marie Candito and Mathieu Constant*	
14:35-15:00	A Wide-Coverage Symbolic Natural Language Inference System *Stergios Chatzikyriakidis and Jean-Philippe Bernardy*	
13:45-15:00	**Parallel session B: Morphology and Syntax** *Chair: Kairit Sirts*	PUB3
13:45-14:10	Ensembles of Neural Morphological Inflection Models *Ilmari Kylliäinen and Miikka Silfverberg*	
14:10-14:35	Nefnir: A high accuracy lemmatizer for Icelandic *Svanhvít Lilja Ingólfsdóttir, Hrafn Loftsson, Jón Friðrik Daðason and Kristín Bjarnadóttir*	
14:35-15:00	Syntax-based identification of light-verb constructions *Silvio Ricardo Cordeiro and Marie Candito*	
15:00-15:30	Coffee Break	

15:30-16:45	**Parallel session A: Machine Learning Applications, Text Classification** *Chair: Jenna Kanerva*	PUB1
15:30-15:55	Natural Language Processing in Policy Evaluation: Extracting Policy Conditions from IMF Loan Agreements *Joakim Åkerström, Adel Daoud and Richard Johansson*	
15:55-16:20	Comparing the Performance of Feature Representations for the Categorization of the Easy-to-Read Variety vs Standard Language *Marina Santini, Benjamin Danielsson and Arne Jönsson*	
16:20-16:45	Unsupervised Inference of Object Affordance from Text Corpora *Michele Persiani and Thomas Hellström*	
15:30-16:45	**Parallel session B: Language Resources and Applications** *Chair: Elena Volodina*	PUB3
15:30-15:55	Annotating evaluative sentences for sentiment analysis: a dataset for Norwegian *Petter Mæhlum, Jeremy Claude Barnes, Lilja Øvrelid and Erik Velldal*	
15:55-16:20	Interconnecting lexical resources and word alignment: How do learners get on with particle verbs? *David Alfter and Johannes Graën*	
16:20-16:45	An Unsupervised Query Rewriting Approach Using N-gram Co-occurrence Statistics to Find Similar Phrases in Large Text Corpora *Hans Moen, Laura-Maria Peltonen, Henry Suhonen, Hanna-Maria Matinolli, Riitta Mieronkoski, Kirsi Telen, Kirsi Terho, Tapio Salakoski and Sanna Salanterä*	
16:45-17:45	**Poster and demo session**	Entrance hall
16:45-17:45	Posters:	
	Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus *Starkaður Barkarson and Steinþór Steingrímsson*	
	May I Check Again? — A simple but efficient way to generate and use contextual dictionaries for Named Entity Recognition. Application to French Legal Texts. *Valentin Barriere and Amaury Fouret*	
	Predicates as Boxes in Bayesian Semantics for Natural Language *Jean-Philippe Bernardy, Rasmus Blanck, Stergios Chatzikyriakidis, Shalom Lappin and Aleksandre Maskharashvili*	
	DIM: The Database of Icelandic Morphology *Kristín Bjarnadóttir, Kristín Ingibjörg Hlynsdóttir and Steinþór Steingrímsson*	

	Bornholmsk Natural Language Processing: Resources and Tools *Leon Derczynski and Alex Speed Kjeldsen*
	Morphosyntactic Disambiguation in an Endangered Language Setting *Jeff Ens, Mika Hämäläinen, Jack Rueter and Philippe Pasquier*
	Tagging a Norwegian Dialect Corpus *Andre Kåsen, Anders Nøklestad, Kristin Hagen and Joel Priestley*
	The Lacunae of Danish Natural Language Processing *Andreas Kirkedal, Barbara Plank, Leon Derczynski and Natalie Schluter*
	Tools for supporting language learning for Sakha *Sardana Ivanova, Anisia Katinskaia and Roman Yangarber*
	Inferring morphological rules from small examples using 0/1 linear programming *Ann Lilliestrom, Koen Claessen and Nicholas Smallbone*
16:45-17:45	Demos:
	LEGATO: A flexible lexicographic annotation tool *David Alfter, Therese Lindström Tiedemann and Elena Volodina*
	The OPUS Resource Repository: An Open Package for Creating Parallel Corpora and Machine Translation Services *Mikko Aulamo and Jörg Tiedemann*
	Garnishing a phonetic dictionary for ASR intake *Iben Nyholm Debess, Sandra Saxov Lamhauge and Peter Juel Henrichsen*
	Docria: Processing and Storing Linguistic Data with Wikipedia *Marcus Klang and Pierre Nugues*
	UniParse: A universal graph-based parsing toolkit *Daniel Varab and Natalie Schluter*
19:30-23:00	Conference Dinner in the Turku Castle

Wednesday October 2, 2019

Time	Session	Room
09:00-09:50	**Keynote by Grzegorz Chrupała: Investigating Neural Representations of Speech and Language** *Chair: Lilja Øvrelid*	PUB1
09:50-10:20	Coffee break	
10:20-12:00	**Parallel session A: Sentiment Analysis and Stance** *Chair: Mathias Creutz*	PUB1
10:20-10:45	Lexicon information in neural sentiment analysis: a multi-task learning approach *Jeremy Claude Barnes, Samia Touileb, Lilja Øvrelid and Erik Velldal*	
10:45-11:10	Aspect-Based Sentiment Analysis using BERT *Mickel Hoang, Oskar Alija Bihorac and Jacobo Rouces*	
11:10-11:35	Political Stance Detection for Danish *Rasmus Lehmann and Leon Derczynski*	
11:35-12:00	Joint Rumour Stance and Veracity Prediction *Anders Edelbo Lillie, Emil Refsgaard Middelboe and Leon Derczynski*	
10:20-12:00	**Parallel session B: Named Entity Recognition** *Chair: Manex Agirrezabal*	PUB3
10:20-10:45	Towards High Accuracy Named Entity Recognition for Icelandic *Svanhvít Lilja Ingólfsdóttir, Sigurjón Þorsteinsson and Hrafn Loftsson*	
10:45-11:10	Named-Entity Recognition for Norwegian *Bjarte Johansen*	
11:10-11:35	Neural Cross-Lingual Transfer and Limited Annotated Data for Named Entity Recognition in Danish *Barbara Plank*	
11:35-12:00	Projecting named entity recognizers without annotated or parallel corpora *Jue Hou, Maximilian Koppatz, José María Hoya Quecedo and Roman Yangarber*	
12:00-13:00	Lunch break	Holiday Club Caribia
13:00-14:00	NEALT business meeting	PUB1
14:00-15:15	**Parallel session A: Text Generation and Language Model Applications** *Chair: Leon Derczynski*	PUB1
14:00-14:25	Template-free Data-to-Text Generation of Finnish Sports News *Jenna Kanerva, Samuel Rönnqvist, Riina Kekki, Tapio Salakoski and Filip Ginter*	

14:25-14:50	Matching Keys and Encrypted Manuscripts *Eva Pettersson and Beata Megyesi*	
14:50-15:15	The Seemingly (Un)systematic Linking Element in Danish *Sidsel Boldsen and Manex Agirrezabal*	
14:00-15:15	**Parallel session B: Speech** *Chair: Grzegorz Chrupała*	PUB3
14:00-14:25	Perceptual and acoustic analysis of voice similarities between parents and young children *Evgeniia Rykova and Stefan Werner*	
14:25-14:50	Enhancing Natural Language Understanding through Cross-Modal Interaction: Meaning Recovery from Acoustically Noisy Speech *Ozge Alacam*	
14:50-15:15	Predicting Prosodic Prominence from Text with Pre-trained Contextualized Word Representations *Aarne Talman, Antti Suni, Hande Celikkanat, Sofoklis Kakouros, Jörg Tiedemann and Martti Vainio*	
15:15-15:45	Coffee Break	
15:45-16:25	**Tutorial on Finnish by *Fred Karlsson*** *Chair: Filip Ginter*	PUB1
16:25-16:35	Closing and announcement of NoDaLiDa'21	PUB1

Abstracts

Keynote: Marie-Catherine de Marneffe

Tuesday October 1, 09:15-10:05 (PUB1)

Do you know that there's still a chance? Identifying speaker commitment for natural language understanding

When we communicate, we infer a lot beyond the literal meaning of the words we hear or read. In particular, our understanding of an utterance depends on assessing the extent to which the speaker stands by the event she describes. An unadorned declarative like "The cancer has spread" conveys firm speaker commitment of the cancer having spread, whereas "There are some indicators that the cancer has spread" imbues the claim with uncertainty. It is not only the absence vs. presence of embedding material that determines whether or not a speaker is committed to the event described: from (1) we will infer that the speaker is committed to there being war, whereas in (2) we will infer the speaker is committed to relocating species not being a panacea, even though the clauses that describe the events in (1) and (2) are both embedded under "(s)he doesn't believe".

(1) The problem, I'm afraid, with my colleague here, he really doesn't believe that it's war.

(2) Transplanting an ecosystem can be risky, as history shows. Hellmann doesn't believe that relocating species threatened by climate change is a panacea.

In this talk, I will first illustrate how looking at pragmatic information of what speakers are committed to can improve NLP applications. Previous work has tried to predict the outcome of contests (such as the Oscars or elections) from tweets. I will show that by distinguishing tweets that convey firm speaker commitment toward a given outcome (e.g., "Dunkirk will win Best Picture in 2018") from ones that only suggest the outcome (e.g., "Dunkirk might have a shot at the 2018 Oscars") or tweets that convey the negation of the event ("Dunkirk is good but not academy level good for the Oscars"), we can outperform previous methods. Second, I will evaluate current models of speaker commitment, using the CommitmentBank, a dataset of naturally occurring discourses developed to deepen our understanding of the factors at play in identifying speaker commitment. We found that a linguistically informed model outperforms a LSTM-based one, suggesting that linguistic knowledge is needed to achieve robust language understanding. Both models however fail to generalize to the diverse linguistic constructions present in natural language, highlighting directions for improvement.

Multilinguality and Machine Translation

Tuesday October 1, 10:35-12:15 (PUB1)

Comparison between NMT and PBSMT Performance for Translating Noisy User-Generated Content
José Carlos Rosales Nunes, Djamé Seddah and Guillaume Wisniewski
This work compares the performances achieved by Phrase-Based Statisti- cal Machine Translation systems (PB- SMT) and attention-based Neuronal Machine Translation systems (NMT) when translating User Generated Con- tent (UGC), as encountered in social medias, from French to English. We show that, contrary to what could be expected, PBSMT outperforms NMT when translating non-canonical inputs. Our error analysis uncovers the speci- ficities of UGC that are problematic for sequential NMT architectures and sug- gests new avenue for improving NMT models.

Bootstrapping UD treebanks for Delexicalized Parsing
Prasanth Kolachina and Aarne Ranta
Standard approaches to treebanking traditionally employ a waterfall model (Sommerville, 2010), where annotation guidelines guide the annotation process and insights from the annotation process in turn lead to subsequent changes in the annotation guidelines. This process remains a very expensive step in creating linguistic resources for a target language, necessitates both linguistic expertise and manual effort to develop the annotations and is subject to inconsistencies in the annotation due to human errors. In this paper, we propose an alternative approach to treebanking—one that requires writing grammars. This approach is motivated specifically in the context of Universal Dependencies, an effort to develop uniform and cross-lingually consistent treebanks across multiple languages. We show here that a bootstrapping approach to treebanking via interlingual grammars is plausible and useful in a process where grammar engineering and treebanking are jointly pursued when creating resources for the target language. We demonstrate the usefulness of synthetic treebanks in the task of delexicalized parsing. Our experiments reveal that simple models for treebank generation are cheaper than human annotated treebanks, especially in the lower ends of the learning curves for delexicalized parsing, which is relevant in particular in the context of low-resource languages.

Lexical Resources for Low-Resource PoS Tagging in Neural Times
Barbara Plank and Sigrid Klerke
More and more evidence is appearing that integrating symbolic lexical knowledge into neural models aids learning. This contrasts the widely-held belief that neural networks largely learn their own feature representations. For example, recent work has shows benefits of integrating lexicons to aid cross-lingual part-of-speech (PoS). However, little is known on how complementary such additional information is, and to what extent improvements depend on the coverage and quality of these external resources. This paper seeks to fill this gap by providing a thorough analysis on the contributions of lexical resources for cross-lingual PoS tagging in neural times.

Toward Multilingual Identification of Online Registers
Veronika Laippala, Roosa Kyllönen, Jesse Egbert, Douglas Biber and Sampo Pyysalo
We consider cross- and multilingual text classification approaches to the identification of online registers (genres), i.e. text varieties with specific situational characteristics. Register is the most important predictor of linguistic variation, and register information could improve the potential of online data for many applications. We introduce the first manually annotated non-English corpus of online registers featuring the full range of linguistic variation found online. The data set consists of 2,237 Finnish documents and follows the register taxonomy developed for the Corpus of Online Registers of English (CORE). Using CORE and the newly introduced corpus, we demonstrate the feasibility of cross-lingual register identification using a simple approach based on convolutional neural networks and multilingual word embeddings. We further find that register identification results can be improved through multilingual training even when a substantial number of annotations is available in the target language.

Embeddings, Biases and Language Change

Tuesday October 1, 10:35-12:15 (PUB3)

Gender Bias in Pretrained Swedish Embeddings
Magnus Sahlgren and Fredrik Olsson
This paper investigates the presence of gender bias in pretrained Swedish embeddings. We focus on a scenario where names are matched with occupations, and we demonstrate how a number of standard pretrained embeddings handle this task. Our experiments show some significant differences between the pretrained embeddings, with word-based methods showing the most bias and contextualized language models showing the least. We also demonstrate that the previously proposed debiasing method does not affect the performance of the various embeddings in this scenario.

A larger-scale evaluation resource of terms and their shift direction for diachronic lexical semantics
Astrid van Aggelen, Antske Fokkens, Laura Hollink and Jacco van Ossenbruggen
Determining how words have changed their meaning is an important topic in Natural Language Processing. However, evaluations of methods to characterise such change have been limited to small, handcrafted resources. We introduce an English evaluation set which is larger, more varied, and more realistic than seen to date, with terms derived from a historical thesaurus. Moreover, the dataset is unique in that it represents change as a shift from the term of interest to a WordNet synset. Using the synset lemmas, we can use this set to evaluate (standard) methods that detect change between word pairs, as well as (adapted) methods that detect the change between a term and a sense overall. We show that performance on the new data set is much lower than earlier reported findings, setting a new standard.

Some steps towards the generation of diachronic WordNets
Yuri Bizzoni, Marius Mosbach, Deitrich Klakow and Stefania Degaetano-Ortlieb
We apply hyperbolic embeddings to trace the dynamics of change of conceptual-semantic relationships in a large diachronic scientific corpus (200 years). Our focus is on emerging scientific fields and the increasingly specialized terminology establishing around them. Reproducing high-quality hierarchical structures such as WordNet on a diachronic scale is a very difficult task. Hyperbolic embeddings can map partial graphs into low dimensional, continuous hierarchical spaces, making more explicit the latent structure of the input. We show that starting from simple lists of word pairs (rather than a list of entities with directional links) it is possible to build diachronic hierarchical semantic spaces which allow us to model a process towards specialization for selected scientific fields.

An evaluation of Czech word embeddings
Karolína Hořeňovská
We present an evaluation of Czech low-dimensional distributed word representations, also known as word embeddings. We describe five different approaches to training the models and three different corpora used in training. We evaluate the resulting models on five different datasets, report the results and provide their further analysis.

Semantics

Tuesday October 1, 13:45-15:00 (PUB1)

Language Modeling with Syntactic and Semantic Representation for Sentence Acceptability Predictions
Adam Ek, Jean-Philippe Bernardy and Shalom Lappin
In this paper, we investigate the effect of enhancing lexical embeddings in LSTM language models (LM) with syntactic and semantic representations. We evaluate the language models using perplexity, and we evaluate the performance of the models on the task of predicting human sentence acceptability judgments. We train LSTM language models on sentences automatically annotated with universal syntactic dependency roles (Nivre, 2016), dependency depth and universal semantic tags (Abzianidze et al., 2017) to predict sentence acceptability judgments. Our experiments indicate that syntactic tags lower perplexity, while semantic tags increase it. Our experiments also show that neither syntactic nor semantic tags improve the performance of LSTM language models on the task of predicting sentence acceptability judgments.

Comparing linear and neural models for competitive MWE identification
Hazem Al Saied, Marie Candito and Mathieu Constant
In this paper, we compare the use of linear versus neural classifiers in a greedy transition system for MWE identification. Both our linear and neural models achieve a new state-of-the-art on the PARSEME 1.1 shared task data sets, comprising 20 languages. Surprisingly, our best model is a simple feed-forward network with one hidden layer, although more sophisticated (recurrent) architectures were tested. The feedback from this study is that tuning a SVM is rather straightforward, whereas tuning our neural system revealed more challenging. Given the number of languages and the variety of linguistic phenomena to handle for the MWE identification task, we have designed an accurate tuning procedure, and we show that hyperparameters are better selected by using a majority-vote within random search configurations rather than a simple best configuration selection. Although the performance is rather good (better than both the best shared task system and the average of the best per-language results), further work is needed to improve the generalization power, especially on unseen MWEs.

A Wide-Coverage Symbolic Natural Language Inference System
Stergios Chatzikyriakidis and Jean-Philippe Bernardy
We present a system for Natural Language Inference which uses a dynamic semantics converter from abstract syntax trees to Coq types. It combines the fine-grainedness of a dynamic semantics system with the powerfulness of a state-of-the-art proof assistant, like Coq. We evaluate the system on all sections of the FraCaS test suite, excluding section 6. This is the first system that does a complete run on the anaphora and ellipsis sections of the FraCaS. It has a better overall accuracy than any previous system.

Morphology and Syntax

Tuesday October 1, 13:45-15:00 (PUB3)

Ensembles of Neural Morphological Inflection Models
Ilmari Kylliäinen and Miikka Silfverberg
We investigate different ensemble learning techniques for neural morphological inflection using bidirectional LSTM encoder-decoder models with attention. We experiment with weighted and unweighted majority voting and bagging. We find that all investigated ensemble methods lead to improved accuracy over a baseline of a single model. However, contrary to expectation based on earlier work by Najafi2018 andSilfverberg2017, weighting does not deliver clear benefits. Bagging was found to underperform plain voting ensembles in general.

Nefnir: A high accuracy lemmatizer for Icelandic
Svanhvít Lilja Ingólfsdóttir, Hrafn Loftsson, Jón Friðrik Daðason and Kristín Bjarnadóttir
Lemmatization, finding the basic morphological form of a word in a corpus, is an important step in many natural language processing tasks when working with morphologically rich languages. We describe and evaluate Nefnir, a new open source lemmatizer for Icelandic. Nefnir uses suffix substitution rules, derived from a large morphological database, to lemmatize tagged text. Evaluation shows that for correctly tagged text, Nefnir obtains an accuracy of 99.55%, and for text tagged with a PoS tagger, the accuracy obtained is 96.88%.

Syntax-based identification of light-verb constructions
Silvio Ricardo Cordeiro and Marie Candito
This paper analyzes results on light-verb construction identification from the PARSEME shared-task, distinguishing between simple cases that could be directly learned from training data from more complex cases that require an extra level of semantic processing.We propose a simple baseline that beats the state of the art for the simple cases, and couple it with another simple baseline to handle the complex cases.We additionally present two other classifiers based on a richer set of features, with results surpassing the state of the art by 8 percentage points.

Machine Learning Applications, Text Classification

Tuesday October 1, 15:30-16:45 (PUB1)

Natural Language Processing in Policy Evaluation: Extracting Policy Conditions from IMF Loan Agreements
Joakim Åkerström, Adel Daoud and Richard Johansson
Social science researchers often use text as the raw data in investigations: for instance, when investigating the effects of IMF policies on the development of countries under IMF programs, researchers typically encode structured descriptions of the programs using a time-consuming manual effort. Making this process automatic may open up new opportunities in scaling up such investigations. As a first step towards automatizing this coding process, we describe an experiment where we apply a sentence classifier that automatically detects mentions of policy conditions in IMF loan agreements and divides them into different types. The results show that the classifier is generally able to detect the policy conditions, although some types are hard to distinguish.

Comparing the Performance of Feature Representations for the Categorization of the Easy-to-Read Variety vs Standard Language
Marina Santini, Benjamin Danielsson and Arne Jönsson
We explore the effectiveness of four feature representations – bag-of-words, word embeddings, principal components and autoencoders – for the binary categorization of the easy-to-read variety vs standard language. Standard language refers to the ordinary language variety used by a population as a whole or by a community, while the “easy-to-read” variety is a simpler (or a simplified) version of the standard language. We test the efficiency of these feature representations on three corpora, which differ in size, class balance, unit of analysis, language and topic. We rely on supervised and unsupervised machine learning algorithms. Results show that bag-of-words is a robust and straightforward feature representation for this task and performs well in many experimental settings. Its performance is equivalent or equal to the performance achieved with principal components and autoencorders, whose preprocessing is however more time-consuming. Word embeddings are less accurate than the other feature representations for this classification task.

Unsupervised Inference of Object Affordance from Text Corpora
Michele Persiani and Thomas Hellström
Affordances denote actions that can be performed in the presence of different objects, or possibility of action in an environment. In robotic systems, affordances and actions may suffer from poor semantic generalization capabilities due to the high amount of required hand-crafted specifications. To alleviate this issue, we propose a method to mine for object-action pairs in free text corpora, successively training and evaluating different prediction models of affordance based on word embeddings.

Language Resources and Applications

Tuesday October 1, 15:30-16:45 (PUB3)

Annotating evaluative sentences for sentiment analysis: a dataset for Norwegian
Petter Mæhlum, Jeremy Claude Barnes, Lilja Øvrelid and Erik Velldal
This paper documents the creation of a large-scale dataset of evaluative sentences – i.e. both subjective and objective sentences that are found to be sentiment-bearing – based on mixed-domain professional reviews from various news-sources. We present both the annotation scheme and first results for classification experiments. The effort represents a step toward creating a Norwegian dataset for fine-grained sentiment analysis.

Interconnecting lexical resources and word alignment: How do learners get on with particle verbs?
David Alfter and Johannes Graën
In this paper, we present a prototype for an online exercise aimed at learners of English and Swedish that serves multiple purposes. The exercise allows learners of the aforementioned languages to train their knowledge of particle verbs receiving clues from the exercise application. The user themselves decide which clue to receive and pay in virtual currency for each, which provides us with valuable information about the utility of the clues that we provide as well as the learners willingness to trade virtual currency versus accuracy of their choice. As resources, we use list with annotated levels from the proficiency scale defined by the Common European Framework of Reference (CEFR) and a multilingual corpus with syntactic dependency relations and word annotation for all language pairs. From the latter resource, we extract translation equivalents for particle verb construction together with a list of parallel corpus examples that can be used as clues in the exercise.

An Unsupervised Query Rewriting Approach Using N-gram Co-occurrence Statistics to Find Similar Phrases in Large Text Corpora

Hans Moen, Laura-Maria Peltonen, Henry Suhonen, Hanna-Maria Matinolli, Riitta Mieronkoski, Kirsi Telen, Kirsi Terho, Tapio Salakoski and Sanna Salanterä

We present our work towards developing a system that should find, in a large text corpus, contiguous phrases expressing similar meaning as a query phrase of arbitrary length. Depending on the use case, this task can be seen as a form of (phrase-level) query rewriting. The suggested approach works in a generative manner, is unsupervised and uses a combination of a semantic word n-gram model, a statistical language model and a document search engine. A central component is a distributional semantic model containing word n-gram vectors (or embeddings) which models semantic similarities between n-grams of different order. As data we use a large corpus of PubMed abstracts. The presented experiment is based on manual evaluation of extracted phrases for arbitrary queries provided by a group of evaluators. The results indicate that the proposed approach is promising and that the use of distributional semantic models trained with uni-, bi- and trigrams seems to work better than a more traditional unigram model.

Posters

Tuesday October 1, 16:45-17:45 (Entrance hall)

Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus
Starkaður Barkarson and Steinþór Steingrímsson
We present ParIce, a new English-Icelandic parallel corpus. This is the first parallel corpus built for the purposes of language technology development and research for Icelandic, although some Icelandic texts can be found in various other multilingual parallel corpora. We map out which Icelandic texts are available for these purposes, collect aligned data and align other bilingual texts we acquired. We describe the alignment process and how we filter the data to weed out noise and bad alignments. In total we collected 43 million Icelandic words in 4.3 million aligned segment pairs, but after filtering, our corpus includes 38.8 million Icelandic words in 3.5 million segment pairs. We estimate that approximately 5% of the corpus data is noise or faulty alignments while more than 50% of the segments we deleted were faulty. We estimate that our filtering process reduced the number of faulty segments in the corpus by more than 60% while only reducing the number of good alignments by approximately 8%.

May I Check Again? — A simple but efficient way to generate and use contextual dictionaries for Named Entity Recognition. Application to French Legal Texts.
Valentin Barriere and Amaury Fouret
In this paper we present a new method to learn a model robust to typos for a Named Entity Recognition task. Our improvement over existing methods helps the model to take into account the context of the sentence inside a justice decision in order to recognize an entity with a typo. We used state-of-the-art models and enriched the last layer of the neural network with high-level information linked with the potential of the word to be a certain type of entity. More precisely, we utilized the similarities between the word and the potential entity candidates the tagged sentence context. The experiments on a dataset of french justice decisions show a reduction of the relative F1-score error of 32%, upgrading the score obtained with the most competitive fine-tuned state-of-the-art system from 94.85% to 96.52%.

Inferring morphological rules from small examples using 0/1 linear programming
Ann Lilliesträm, Koen Claessen and Nicholas Smallbone
We show how to express the problem of finding an optimal morpheme segmentation from a set of labelled words as a 0/1 linear programming problem, and how to build on this to analyse a language's morphology. The approach works even when there is very little training data available.

Predicates as Boxes in Bayesian Semantics for Natural Language
Jean-Philippe Bernardy, Rasmus Blanck, Stergios Chatzikyriakidis, Shalom Lappin and Aleksandre Maskharashvili
In this paper, we present a Bayesian approach to natural language semantics. Our main focus is on the inference task in an environment where judgments require probabilistic reasoning. We treat nouns, verbs, adjectives, etc. as unary predicates, and we model them as boxes in a bounded domain. We apply Bayesian learning to satisfy constraints expressed as premises. In this way we construct a model, by specifying boxes for the predicates. The probability of the hypothesis (the conclusion) is evaluated against the model that incorporates the premises as constraints.

DIM: The Database of Icelandic Morphology
Kristín Bjarnadóttir, Kristín Ingibjörg Hlynsdóttir and Steinþór Steingrímsson
The topic of this paper is The Database of Icelandic Morphology (DIM), a multi-purpose linguistic resource, created for use in language technology, as a reference for the general public in Iceland, and for use in research on the Icelandic language. DIM contains inflectional paradigms and analysis of word formation, with a vocabulary of approx. 285,000 lemmas. DIM is based on The Database of Modern Icelandic Inflection, which has been in use since 2004.

Bornholmsk Natural Language Processing: Resources and Tools
Leon Derczynski and Alex Speed Kjeldsen
This paper introduces language processing resources and tools for Bornholmsk, a language spoken on the island of Bornholm, with roots in Danish and closely related to Scanian. This presents an overview of the language and available data, and the first NLP models for this living, minority Nordic language.

Morphosyntactic Disambiguation in an Endangered Language Setting
Jeff Ens, Mika Hämäläinen, Jack Rueter and Philippe Pasquier
Endangered Uralic languages present a high variety of inflectional forms in their morphology. This results in a high number of homonyms in inflections, which introduces a lot of morphological ambiguity in sentences. Previous research has employed constraint grammars to address this problem, however CGs are often unable to fully disambiguate a sentence, and their development is labour intensive. We present an LSTM based model for automatically ranking morphological readings of sentences based on their quality. This ranking can be used to evaluate the existing CG disambiguators or to directly morphologically disambiguate sentences. Our approach works on a morphological abstraction and it can be trained with a very small dataset.

Tagging a Norwegian Dialect Corpus
Andre Kåsen, Anders Nøklestad, Kristin Hagen and Joel Priestley
This paper describes an evaluation of five data-driven part-of-speech (PoS) taggers for spoken Norwegian. The taggers all rely on different machine learning mechanisms: decision trees, hidden Markov models (HMMs), conditional random fields (CRFs), long-short term memory networks (LSTMs), and convolutional neural networks (CNNs). We go into some of the challenges posed by the task of tagging spoken, as opposed to written, language, and in particular a wide range of dialects as is found in the recordings of the LIA (Language Infrastructure made Accessible) project. The results show that the taggers based on either conditional random fields or neural networks perform much better than the rest, with the LSTM tagger getting the highest score.

The Lacunae of Danish Natural Language Processing
Andreas Kirkedal, Barbara Plank, Leon Derczynski and Natalie Schluter
Danish is a North Germanic language spoken principally in Denmark, a country with a long tradition of technological and scientific innovation. However, the language has received relatively little attention from a technological perspective. In this paper, we review Natural Language Processing (NLP) research, digital resources and tools which have been developed for Danish. We find that availability of models and tools is limited, which calls for work that lifts Danish NLP a step closer to the privileged languages.

Tools for supporting language learning for Sakha
Sardana Ivanova, Anisia Katinskaia and Roman Yangarber
This paper presents an overview of the available linguistic resources for the Sakha language, and presents new tools for supporting language learning for Sakha. The essential resources include a morphological analyzer, digital dictionaries, and corpora of Sakha texts. Based on these resources, we implement a language-learning environment for Sakha in the Revita CALL platform. We extended an earlier, preliminary version of the morphological analyzer/transducer, built on the Apertium finite-state platform. The analyzer currently has an adequate level of coverage, between 86% and 89% on two Sakha corpora. Revita is a freely available online language learning platform for learners beyond the beginner level. We describe the tools for Sakha currently integrated into the Revita platform. To the best of our knowledge, at present, this is the first large-scale project undertaken to support intermediate-advanced learners of a minority Siberian language.

Demos

Tuesday October 1, 16:45-17:45 (Entrance hall)

LEGATO: A flexible lexicographic annotation tool
David Alfter, Therese Lindström Tiedemann and Elena Volodina
This article is a report from an ongoing project aiming at analyzing lexical and grammatical competences of Swedish as a Second language (L2). To facilitate lexical analysis, we need access to metalinguistic information about relevant vocabulary that L2 learners can use and understand. The focus of the current article is on the lexical annotation of the vocabulary scope for a range of lexicographical aspects, such as morphological analysis, valency, types of multiword units, etc. We perform parts of the analysis automatically, and other parts manually. The rationale behind this is that where there is no possibility to add information automatically, manual effort needs to be added. To facilitate the latter, a tool LEGATO has been designed, implemented and currently put to active testing.

Docria: Processing and Storing Linguistic Data with Wikipedia
Marcus Klang and Pierre Nugues
The availability of user-generated content has increased significantly over time. Wikipedia is one example of a corpora which spans a huge range of topics and is freely available. Storing and processing these corpora requires flexible documents models as they may contain malicious and incorrect data. Docria is a library which attempts to address this issue by providing a solution which can be used with small to large corpora, from laptops using Python interactively in a Jupyter notebook to clusters running map-reduce frameworks with optimized compiled code. Docria is available as open-source code.

UniParse: A universal graph-based parsing toolkit
Daniel Varab and Natalie Schluter
This paper describes the design and use of the graph-based parsing framework and toolkit UniParse, released as an open-source python software package. UniParse as a framework novelly streamlines research prototyping, development and evaluation of graph-based dependency parsing architectures. UniParse does this by enabling highly efficient, sufficiently independent, easily readable, and easily extensible implementations for all dependency parser components. We distribute the toolkit with ready-made configurations as re-implementations of all current state-of-the-art first-order graph-based parsers, including even more efficient Cython implementations of both encoders and decoders, as well as the required specialised loss functions.

The OPUS Resource Repository: An Open Package for Creating Parallel Corpora and Machine Translation Services
Mikko Aulamo and Jörg Tiedemann
This paper presents a flexible and powerful system for creating parallel corpora and for running neural machine translation services. Our package provides a scalable data repository backend that offers transparent data pre-processing pipelines and automatic alignment procedures that facilitate the compilation of extensive parallel data sets from a variety of sources. Moreover, we develop a web-based interface that constitutes an intuitive frontend for end-users of the platform. The whole system can easily be distributed over virtual machines and implements a sophisticated permission system with secure connections and a flexible database for storing arbitrary metadata. Furthermore, we also provide an interface for neural machine translation that can run as a service on virtual machines, which also incorporates a connection to the data repository software.

Garnishing a phonetic dictionary for ASR intake
Iben Nyholm Debess, Sandra Saxov Lamhauge and Peter Juel Henrichsen
We present a new method for preparing a lexical-phonetic database as a resource for acoustic model training. The research is an offshoot of the ongoing Project Ravnur (Speech Recognition for Faroese), but the method is language-independent. At NODALIDA 2019 we demonstrate the method (called SHARP) online, showing how a traditional lexical-phonetic dictionary (with a very rich phone inventory) is transformed into an ASR-friendly database (with reduced phonetics, preventing data sparseness). The mapping procedure is informed by a corpus of speech transcripts. We conclude with a discussion on the benefits of a well-thought-out BLARK design (Basic Language Resource Kit), making tools like SHARP possible.

Keynote: Grzegorz Chrupała

Wednesday October 2, 09:00-09:50 (PUB1)

Investigating Neural Representations of Speech and Language

Learning to communicate in natural language in one of the unique human abilities which are at the same time extraordinarily important and extraordinarily difficult to reproduce in silico. Substantial progress has been achieved in some specific data-rich and constrained cases such as automatic speech recognition or machine translation. However the general problem of learning to use natural language with weak and noisy supervision in a grounded setting is still open. In this talk I will present recent work which addresses this challenge using deep recurrent neural network models. I will then focus on analytical methods which allow us to better understand the nature and localization of representations emerging in such architectures.

Sentiment Analysis and Stance

Wednesday October 2, 10:20-12:00 (PUB1)

Lexicon information in neural sentiment analysis: a multi-task learning approach
Jeremy Claude Barnes, Samia Touileb, Lilja Øvrelid and Erik Velldal
This paper explores the use of multi-task learning (MTL) for incorporating external knowledge in neural models. Specifically, we show how MTL can enable a BiLSTM sentiment classifier to incorporate information from sentiment lexicons. Our MTL set-up is shown to improve model performance (compared to a single-task set-up) on both English and Norwegian sentence-level sentiment datasets. The paper also introduces a new sentiment lexicon for Norwegian.

Aspect-Based Sentiment Analysis using BERT
Mickel Hoang, Oskar Alija Bihorac and Jacobo Rouces
Sentiment analysis has become very popular in both research and business due to the increasing amount of opinionated text from Internet users. Standard sentiment analysis deals with classifying the overall sentiment of a text, but this doesn't include other important information such as towards which entity, topic or aspect within the text the sentiment is directed. Aspect-based sentiment analysis (ABSA) is a more complex task that consists in identifying both sentiments and aspects. This paper shows the potential of using the contextual word representations from the pre-trained language model BERT, together with a fine-tuning method with additional generated text, in order to solve out-of-domain ABSA and outperform previous state-of-the-art results on SemEval-2015 Task 12 subtask 2 and SemEval-2016 Task 5. To the best of our knowledge, no other existing work has been done on out-of-domain ABSA for aspect classification.

Political Stance Detection for Danish
Rasmus Lehmann and Leon Derczynski
The task of stance detection consists of classifying the opinion within a text towards some target. This paper seeks to generate a dataset of quotes from Danish politicians, label this dataset to allow the task of stance detection to be performed, and present annotation guidelines to allow further expansion of the generated dataset. Furthermore, three models based on an LSTM architecture are designed, implemented and optimized to perform the task of stance detection for the generated dataset. Experiments are performed using conditionality and bi-directionality for these models, and using either singular word embeddings or averaged word embeddings for an entire quote, to determine the optimal model design. The simplest model design, applying neither conditionality or bi-directionality, and averaged word embeddings across quotes, yields the strongest results. Furthermore, it was found that inclusion of the quotes politician, and the party affiliation of the quoted politician, greatly improved performance of the strongest model.

Joint Rumour Stance and Veracity Prediction
Anders Edelbo Lillie, Emil Refsgaard Middelboe and Leon Derczynski
The net is rife with rumours that spread through microblogs and social media. Not all the claims in these can be verified. However, recent work has shown that the stances alone that commenters take toward claims can be sufficiently good indicators of claim veracity, using e.g. an HMM that takes conversational stance sequences as the only input. Existing results are monolingual (English) and mono-platform (Twitter). This paper introduces a stance-annotated Reddit dataset for the Danish language, and describes various implementations of stance classification models. Of these, a Linear SVM provides predicts stance best, with 0.76 accuracy / 0.42 macro F1. Stance labels are then used to predict veracity across platforms and also across languages, training on conversations held in one language and using the model on conversations held in another. In our experiments, monolinugal scores reach stance-based veracity accuracy of 0.83 (F1 0.68); applying the model across languages predicts veracity of claims with an accuracy of 0.82 (F1 0.67). This demonstrates the surprising and powerful viability of transferring stance-based veracity prediction across languages.

Named Entity Recognition

Wednesday October 2, 10:20-12:00 (PUB3)

Towards High Accuracy Named Entity Recognition for Icelandic
Svanhvít Lilja Ingólfsdóttir, Sigurjón Þorsteinsson and Hrafn Loftsson
We report on work in progress which consists of annotating an Icelandic corpus for named entities (NEs) and using it for training a named entity recognizer based on a Bidirectional Long Short-Term Memory model. Currently, we have annotated 7,538 NEs appearing in the first 200,000 tokens of a 1 million token corpus, MIM-GOLD, originally developed for serving as a gold standard for part-of-speech tagging. Our best performing model, trained on this subset of MIM-GOLD, and enriched with external word embeddings, obtains an overall F1 score of 81.3% when categorizing NEs into the following four categories: persons, locations, organizations and miscellaneous. Our preliminary results are promising, especially given the fact that 80% of MIM-GOLD has not yet been used for training.

Named-Entity Recognition for Norwegian
Bjarte Johansen
NER is the task of recognizing and demarcating the segments of a document that are part of a name and which type of name it is. We use 4 different categories of names: Locations (LOC), miscellaneous (MISC), organizations (ORG), and persons (PER). Even though we employ state of the art methods—including sub-word embeddings—that work well for English, we are unable to reproduce the same success for the Norwegian written forms. However, our model performs better than any previous research on Norwegian text. The study also presents the first NER for Nynorsk. Lastly, we find that by combining Nynorsk and Bokmål into one training corpus we improve the performance of our model on both languages.

Neural Cross-Lingual Transfer and Limited Annotated Data for Named Entity Recognition in Danish
Barbara Plank
Named Entity Recognition (NER) has greatly advanced by the introduction of deep neural architectures. However, the success of these methods depends on large amounts of training data. The scarcity of publicly-available human-labeled datasets has resulted in limited evaluation of existing NER systems, as is the case for Danish. This paper studies the effectiveness of cross-lingual transfer for Danish, evaluates its complementarity to limited gold data, and sheds light on performance of Danish NER.

Projecting named entity recognizers without annotated or parallel corpora
Jue Hou, Maximilian Koppatz, José María Hoya Quecedo and Roman Yangarber
Named entity recognition (NER) is a well-researched task in the field of NLP, which typically requires large annotated corpora for training usable models. This is a problem for languages which lack large annotated corpora, such as Finnish. We propose an approach to create a named entity recognizer with no annotated or parallel documents, by leveraging strong NER models that exist for English. We automatically gather a large amount of chronologically matched data in two languages, then project named entity annotations from the English documents onto the Finnish ones, by resolving the matches with limited linguistic rules. We use this "artificially" annotated data to train a BiLSTM-CRF model. Our results show that this method can produce annotated instances with high precision, and the resulting model achieves state-of-the-art performance.

Text Generation and Language Model Applications

Wednesday October 2, 14:00-15:15 (PUB1)

Template-free Data-to-Text Generation of Finnish Sports News
Jenna Kanerva, Samuel Rönnqvist, Riina Kekki, Tapio Salakoski and Filip Ginter
News articles such as sports game reports are often thought to closely follow the underlying game statistics, but in practice they contain a notable amount of background knowledge, interpretation, insight into the game, and quotes that are not present in the official statistics. This poses a challenge for automated data-to-text news generation with real-world news corpora as training data. We report on the development of a corpus of Finnish ice hockey news, edited to be suitable for training of end-to-end news generation methods, as well as demonstrate generation of text, which was judged by journalists to be relatively close to a viable product. The new dataset and system source code are available for research purposes.

Matching Keys and Encrypted Manuscripts
Eva Pettersson and Beata Megyesi.
Historical cryptology is the study of historical encrypted messages aiming at their decryption by analyzing the mathematical, linguistic and other coding patterns and their historical context. In libraries and archives we can find quite a lot of ciphers, as well as keys describing the method used to transform the plaintext message into a ciphertext. In this paper, we present work on automatically mapping keys to ciphers to reconstruct the original plaintext message, and use language models generated from historical texts to guess the underlying plaintext language.

The Seemingly (Un)systematic Linking Element in Danish
Sidsel Boldsen and Manex Agirrezabal
The use of a linking element between compound members is a common phenomenon in Germanic languages. Still, the exact use and conditioning of such elements is a disputed topic in linguistics. In this paper we address the issue of predicting the use of linking elements in Danish. Following previous research that shows how the choice of linking element might be conditioned by phonology, we frame the problem as a language modeling task: Considering the linking elements -s/-□ the problem becomes predicting what is most probable to encounter next, a syllable boundary or the joining element, 's'. We show that training a language model on this task reaches an accuracy of 94 %, and in the case of an unsupervised model, the accuracy reaches 80%.

Speech

Wednesday October 2, 14:00-15:15 (PUB3)

Perceptual and acoustic analysis of voice similarities between parents and young children
Evgeniia Rykova and Stefan Werner
Human voice provides the means for verbal communication and forms a part of personal identity. Due to genetic and environmental factors, a voice of a child should resemble the voice of her parent(s), but voice similarities between parents and young children are underresearched. Read-aloud speech of Finnish-speaking and Russian-speaking parent-child pairs was subject to perceptual and multi-step instrumental and statistical analysis. Finnish-speaking listeners could not discriminate family pairs auditorily in an XAB paradigm, but the Russian-speaking listeners' mean accuracy of answers reached 72.5%. On average, in both language groups family-internal f0 similarities were stronger than family-external, with parents showing greater family-internal similarities than children. Auditory similarities did not reflect acoustic similarities in a straightforward way.

Enhancing Natural Language Understanding through Cross-Modal Interaction: Meaning Recovery from Acoustically Noisy Speech
Ozge Alacam
Cross-modality between vision and language is a key component for effective and efficient communication, and human language processing mechanism successfully integrates information from various modalities to extract the intended meaning. However, incomplete linguistic input, i.e. due to a noisy environment, is one of the challenges for a successful communication. In that case, an incompleteness in one channel can be compensated by information from another one. In this paper, by conducting visual-world paradigm, we investigated the dynamics between syntactically possible gap fillers and the visual arrangements in incomplete German sentences and their effect on overall sentence interpretation.

Predicting Prosodic Prominence from Text with Pre-trained Contextualized Word Representations
Aarne Talman, Antti Suni, Hande Celikkanat, Sofoklis Kakouros, Jörg Tiedemann and Martti Vainio
In this paper we introduce a new natural language processing dataset and benchmark for predicting prosodic prominence from written text. To our knowledge this will be the largest publicly available dataset with prosodic labels. We describe the dataset construction and the resulting benchmark dataset in detail and train a number of different models ranging from feature-based classifiers to neural network systems for the prediction of discretized prosodic prominence. We show that pre-trained contextualized word representations from BERT outperform the other models even with less than 10% of the training data. Finally

we discuss the dataset in light of the results and point to future research and plans for further improving both the dataset and methods of predicting prosodic prominence from text. The dataset and the code for the models will be made publicly available.

Tutorial on Finnish: Fred Karlsson

Wednesday October 2, 15:45-16:25 (PUB1)

Fred Karlsson is a Professor Emeritus of general linguistics at the University of Helsinki. Although he is a Swedish-speaking Finn, Fred's knowledge of his second native language, Finnish, is exceptionally good, and he is widely considered a de facto authority on the language's rules. Fred is also a renowned computational linguist, having designed the *Constraint Grammar* language-independent formalism for automatic morphological disambiguation and syntactic analysis. He has also worked on the history of linguistics, co-authoring *The History of Linguistics in the Nordic Countries*, among other contributions.

Local information

City map

Scan or click →

SCAN ME

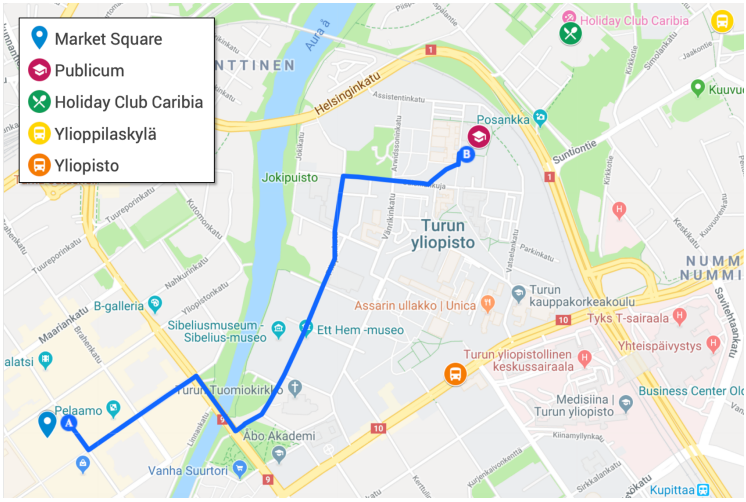
Conference venue

All conference programs (registration, main conference sessions, workshops, poster/demo sessions and the NEALT business meeting) will take place in the **Publicum** building of the University of Turku. The building is located in the university campus area, in walking distance from the city centre.

- **Address:** Publicum, Assistentinkatu 7, 20500 Turku

SCAN ME

It takes about 20 minutes to walk there from the **Market Square** (**Kauppatori**), city center. If you want to take the bus, use platform **T42** and take bus number **50**, **51**, **53** or **54** and get off at their last bus stop which is **Ylioppilaskylä** (1644), close to Holiday Club Caribia. From there it takes about 5 minutes to walk to the Publicum building (see the next page). Alternatively, you can get off at bus stop **Yliopisto** (115). The walk from here takes about 10 minutes.



You can use the ”JOURNEY PLANNER” at the local bus company’s website (`https://www.foli.fi/en`) to find busses, bus stops, time tables etc.

Lunch breaks

Lunches are served in restaurant Terrace in **Holiday Club Caribia**, which is within 5 minutes walking distance from the conference venue.

- **Address:** Kongressikuja 1, 20540 Turku

SCAN ME

Welcome reception

The welcome reception will take place in **Old Town Hall**. This event is sponsored by the city of Turku.

Image: Wikipedia

- **Date/Time:** Monday 30.9.2019, 19.00-20.30
- **Address:** Old Town Hall, Aurakatu 2, Turku
- In the city centre, 20-30 min walk from the conference, 2km
- Snacks and a drink

SCAN ME

Conference dinner

The conference dinner will be held in the **Turku Castle**.

Image: Visit Turku / Museokeskus

SCAN ME

- **Date/Time:** Tuesday 1.10.2019, 19.30-23.00
- **Address:** Turku Castle, King's Hall, Linnankatu 80
- **Full dinner and drinks**
- **Transport there**:
  - 30-45 min (3km) walk from the city centre (Kauppatori), following the Aura river towards the harbor.
  - By bus: Use the platform **T89** in the city centre and take bus number **1** and get off at bus stop **Turun linna** / **Turku Castle** (54). There is a bus leaving approximately every seven minute and the ride takes about 15 minutes. (visit: `www.foli.fi/en/central-bus-stops-turku`)
  - Rental city bikes and rental kickbikes are other transportation options (visit: `https://www.foli.fi/en/node/4543/`)
- **Transport back:** the above options, plus **we will have several additional busses leaving the venue** at **23.10**, stopping in the centre and by Holiday Club Caribia. If you miss these, there is still a city bus leaving at 23.20.
  - Direction either "Keskusta" (centre) or "Lentoasema" (airport)
  - Timetables for bus number 1
    * 20:00 20:07 20:15 20:22 20:30 20:37 20:45 20:52
    * 21:00 21:15 21:30 21:45
    * 22:00 22:20 22:50
    * 23:20

Transportation

Turku city centre (bus stops: *Keskusta* or *Kauppatori*)

The city centre (*Keskusta* in Finnish) is the main hub for all buses going to different directions in the city. Some of the buses going towards the city have *Kauppatori* (lit. Market Square) as their terminal station. For a map of all bus platforms in the city centre, please visit `https://www.foli.fi/en/central-bus-stops-turku`. Please note that street signs in Turku have the street names in Finnish and Swedish.

Different transportation means are available:

- **Bus:** Busses commute regularly in the city. There are also occasional night busses, operating after the midnight. You can use the "JOURNEY PLANNER" in bus company website (`https://www.foli.fi/en`) for searching bus stops, busses, time tables and the local map.
- **Taxi:** You can book a taxi from the number 0210041 in Finland and +358 600 14121 from abroad. Cost of the call is 1,84€ + telecom operator fees. visit: `https://taxidata.fi/?lang=en`
- **City-bike:** Rental city bikes and rental kickbikes are other transport options
  visit: `https://www.foli.fi/en/node/4543/`

Turku Airport to/from city centre

If you are in the airport, you can use bus number **1** (in front of main entrance/exit door) to get to the city centre, in about 20 minutes. There are also taxis, pretty close to the main door. To reach the airport from city centre, use platform **T86** and take bus number **1**.

Free WIFI networks at the venue

In the Publicum building (conference venue), two free WIFI networks are available:

1. **eduroam:** If you have previously used the eduroam network in your home institute, you should be able to use the same username and password to connect to all eduroam hotspots at the University of Turku and use the internet.

2. **UTU Visitor:** In case you do not have an eduroam account, you can use the "UTU Visitor" network. To begin using the network, choose UTU Visitor from the list of available networks on your device. After you have connected to the network, you will need to open a browser window and you will be redirected to the registration page.

   On the registration page, enter your name and an email address (where the login credentials will be sent), and accept the network's terms and conditions. You will be logged in automatically and redirected to the network's guide service.

   - UTU Visitor allows only normal browser traffic, tunneling protocols, and secure email protocols.
   - Your login credentials are active for a fixed period of time. A single user can connect a maximum of two devices to the network.

In case of emergency

In case of emergency and if you need a medicine, you can always reach to the **Yliopiston Apteekki**, the main drugstore of the city, located in city centre area. This drugstore is usually open until 23:00.

Address: Yliopistonkatu 25, 20100 Turku , **Phone:** +358 300 20200.

If you really need to make an emergency call, you can dial *112* in Finland. The police, rescue services, patient transportation and social services are all available in urgent need for help.